DOCUMENT RESUME

ED 053 175 TM 000 702

AUTHOR Klein, Stephen P.
TITLE The Uses and Limitations of Standardized Tests in Meeting the Demands for Accountability.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
PUB DATE Jan 71
NOTE 20p.; UCLA Evaluation Comment, v2 n4

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement, Direction Writing, *Educational Accountability, Educational Objectives, Evaluation, *Evaluation Needs, Formative Evaluation, Instructional Improvement, Instructional Programs, *Performance Contracts, Performance Criteria, Program Planning, Scores, *Standardized Tests, Test Construction, Testing, *Test Validity

ABSTRACT

Major demands of accountability as related to performance contracting are outlined, as well as the limitations of standardized tests in meeting these demands. (AG)

UCLA CSE Evaluation comment

January, 1971 Vol. 2, No. 4

Center for the Study of Evaluation

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED053175

TM 000 702

THE USES AND LIMITATIONS OF STANDARDIZED TESTS IN MEETING THE DEMANDS FOR ACCOUNTABILITY[1]

Stephen P. Klein

University of California, Los Angeles

The two major demands of "accountability" as it relates to performance contracting are to provide a *valid* system of assessing student performance and to provide a *fair* system for paying the contractor. The four major limitations of existing standardized tests in meeting the first of these two demands are (1) likelihood of poor *overlap* between the test's and the school's objectives and the priorities associated with these objectives, (2) inappropriate *test designs and formats* for the target populations, (3) difficult and confusing *test instructions and administration procedures* that introduce irrelevant factors into a student's score, and (4) low *test validity* in the sense that the tests do not really assess the kinds of student skills and abilities that their titles imply they do.

Among the important implications of these limitations are (1) reduction of the value of standardized tests as a basis for a fair payment system, (2) alienation of educators from the principles of accountability, since educators are responsible for improving student performance on irrelevant measures and, finally, (3) reduction of the test's sensitivity to the point that it is almost impossible to identify just which educational programs are really making positive contributions. The nature of most performance contracts and their reliance on standardized test scores and average grade-norm changes further reduces the effectiveness of applying accountability principles. The reason for this is that such contracts often fail to handle important technical problems associated with measuring actual and relevant gains in student performance.

The solution to these problems does not lie solely in finding better ways to use existing measures. It must also include improvements in test formats, instructions, content, and administration procedures, as well as in methods for interpreting test results.

[1]*Based on an invited address to the AASA National Academy for School Executives; Dallas, Texas; October 6, 1970. The author wishes to express his appreciation to members of the Center staff and to W. Stanley Kruger for their comments and suggestions on the paper.*

"Why can't Johnny read?" was the complaint of the 1960's. Today, federal and state governments, school officials, and parents are asking, "Who is responsible for the situation?", "How can we remedy it?", and "How can we be sure that it is corrected?" These questions, especially the last one, have caused an increased interest in evaluation and have introduced terms like "accountability" and "performance contracting" to educational jargon. Although these terms may be defined in several ways, they may be generally characterized in terms of two or three basic premises. First, educators should be responsible, or "accountable," for producing actual gains in student achievement. That they may be certified professionals who go through the motions of teaching will not suffice; they must produce measurable gains in student achievement. Further, these gains must occur on the desired objectives which have been specified in advance rather than on some nebulous objectives possibly chosen after the fact. This demand is parallel to the one you make when you go to the dentist; you want him to fix the tooth that hurts, not the one next to it.

"Performance contracting" is associated with "accountability" since it provides one approach to trying to achieve and measure the desired objectives in addition to holding educators responsible for them. Briefly, a "performance contract" is a written agrement between two parties (such as a school district and a private firm, or a principal and a teacher) in which one of the parties agrees to produce certain changes in student performance. These contracts are usually written in terms of the teacher or private firm agreeing to improve student achievement in prespecified areas, such as reading comprehension, in return for some reward or payment from the school. The size of this remuneration is often a direct function of the amount of improvement obtained as indicated by some previously agreed-upon index, such as a standardized test score. For example, Westinghouse Learning Corporation has performance contracts with school districts in California, New Mexico, Nevada, and Pennsylvania to improve the reading and mathematics skills of disadvantaged children.

The districts have agreed to pay Westinghouse for providing instruction only in terms of the degree to which the

desired changes in student performance have been produced. In other words, the better the students' performance, the more money Westinghouse gets. Behavioral Research Laboratories of Palo Alto even has a contract to run an entire school and which provides a money-back guarantee that it will produce the desired gains in student performance (*U.S. News & World Report*, 1970).

THE TWO DEMANDS

One key feature of "accountability" and the practice of writing performance contracts is the need for accurate assessments of student performance. It is no longer sufficient for a teacher or contractor to say, "The students really enjoyed the course and I think they learned a great deal." Today, claims such as these must be supported by concrete and valid evidence. This need has led to the widespread use of standardized tests in performance contracts. One Office of Economic Opportunity study, for example, involves six performance contractors in 18 school districts in grades 1 to 3 and 7 to 9 with over 4,000 pupils (OEO, 1970). An important component of this study is the stipulation that at least three different standardized tests be used in each classroom at the beginning and end of the year. Thus, in this study, it is apparently assumed that standardized tests will meet the two basic demands of accountability. The first of these two demands is to provide a *valid system* for assessing relevant student performance. Where performance contracts are used to achieve accountability, the second demand is to provide a *fair system* for paying the contractor in terms of the degree to which he influenced this performance.

Let us consider the implications of using tests in terms of trying to meet each of these two demands. The first demand, providing a valid system for assessing student performance, assumes that good methods are available for measuring pupil performance on all the goals a school might like to achieve. This assumption is necessary to all accountability systems and especially to those involving performance contracts, since there must be a legally acceptable method of determining whether the "contractor" did, indeed, achieve the desired gains in those areas of student performance of primary interest to the school and the community. That such is not the case, however, is indicated even by a cursory inspection of the measurement armory. First, certain kinds of assessment methods (such as observations, teacher ratings, or essay tests) are generally not sufficiently valid or reliable for accountability demands. Second, though objective tests (standardized or otherwise) are readily available for some goal areas like "reading comprehension," for others like "the desire to read" they are not readily available. This situation has led to the practice of holding responsible parties "accountable" only for those goal areas which are amenable to measurement with existing standardized instruments while omitting from the contract any specifications for performance on goals that are difficult to assess — a classic case of the tail wagging the dog. As students, we were rewarded for studying what we thought a test would cover and not necessarily what we thought were the most important or interesting aspects of a course. Similarly, when an educator or contractor is held accountable for performance in one area and not another, it is obvious where his principal efforts will be devoted. It is apparent, therefore, that in current performance measurement the reliance on existing instruments may actually detract from the total set of goals the school is trying to achieve. Thus, the expedient procedure of relying on available standardized measures to assess some objectives while ignoring others may lead to long-term failure of a school's educational program.

OVERLAP

Even if we grant that the areas covered by existing instruments are those in which we are most interested (which appears *not* to be the case), we must still accurately measure student performance in these areas. A firm which contracts to improve student performance in mathematics, for example, will certainly not agree to have its efforts judged by a reading test. Similarly, a school which is letting the contract will want a mathematics test covering the particular objectives with which it is most concerned and not some other set. This accountability principle of appropriate assessment procedures in conjunction with the nature of business contracts has led to the practice of specifying in advance just which standardized measures will be used to judge student progress. One problem with this approach is that it is highly improbable that any test will *overlap* well the particular set of objectives a given school program is trying to achieve. Faced with this problem, schools generally use the one or two instruments that will result in the best possible overlap between test and program objectives. This would be a reasonable compromise were it not for the fact that most tests usually provide only a single score summed over all the objectives they measure. This may lead to far less precise assessment than the level needed for true accountability. To illustrate this point, suppose a school has a science program with ten major objectives. Search of test catalogues and other sources indicates that the measure that best overlaps these objectives is the "XYZ" Science Test. This test yields only a single score and covers eight of the objectives, but it also covers four additional objectives in which the school is not at all interested, such as memory for plant names. Further, student performance on some of these extraneous objectives may be improved more easily than on those of central interest. Thus, instead of getting what appears to be 80% overlap (8 out of 10), the school gets only 67% (8 out of 12). In fact, if the contractor focuses on some of the extraneous objectives in order to boost his rewards for student progress, it may be even less than 67%.

The foregoing example further assumes that the relative importance of the eight overlapping objectives is the same in both the program and the test. Obviously, if an existing test focuses more on some objectives than on others, then this differential may not coincide well with the relative values assigned to objectives in a given program. This kind of problem will further reduce the effective percentage of overlap between the test's and the program's objectives.

The manual for a popular mathematics test, for example, claims that the instrument measures student performance in eight separate areas. On closer inspection, however, it is apparent that 80% of the test's total items are accounted for by only four of these areas. An examination of the average item difficulties presented in the manual also reveals that two of these four areas contain much easier items than the other two. In short, how well a student does on this test is not primarily a function of his mastery in eight areas, but rather in only four of them; and his overall score indicates the extent to which he can answer easy items in some areas and difficult ones in another rather than how well he can perform in each of the eight. Further, a review of popular standardized instruments and their manuals indicates that this problem is not specific to just one or two measures, but is typical of the field (CSE, 1970).

A second approach to solving the problem of overlap has been to specify several measures in the performance contract to ensure the assessment of as many objectives as possible. This is not an especially good approach since it still includes inappropriate items for the particular objec-

tives of a given program and allows low correspondence between test and program priorities on these objectives. Further, it causes higher testing costs, cumbersome testing methods and, frequently, longer testing times. It appears, therefore, that the solution to the overlap problem does not lie in finding the best combination of existing tests, but rather in developing a new approach to test construction, test scoring, and test interpretation (Klein, 1970).

So far, then, one important limitation of the use of standardized tests for accountability is indicated in the *poor overlap* that frequently exists between test and school objectives. This problem is principally caused by (1) the failure of standardized measures to cover many important goal areas and the probable exclusion of these areas from performance contracts which, in turn, may result in their effective exclusion from a school's educational program, and (2) the likelihood of poor correspondence between a test's and a program's objectives and the relative importance attached to them. It was also noted that the latter situation may result in the contractor's emphasizing a given *test's* objectives and priorities rather than the *school's* objectives since his payment is based on changes in student scores on the whole test and not just those parts germane to the particular needs of the pupils. A second and more subtle consequence of poor overlap between test and program objectives is that the measuring instruments used may not be as sensitive to desired changes in student performance as they should be. Thus, though performance may or may not improve or, perhaps, may even regress on key objectives, these changes may be hidden because of the conglomerate nature of a test's total score. This generally works to the disadvantage of both the contractor and the school since it will make it more difficult both to produce score changes and to know what kinds of performance changes have and have not occurred.

This problem of sensitivity of measurement to relevant objectives is analogous to assessing whether federal programs to fight air pollution are working. If the average amount of daily pollution in the United States is used as a criterion, then it is unlikely that any significant changes will be registered over a period of several years. If, on the other hand, the levels of air pollution were recorded in those urban areas where different programs were being tried, then which of these programs were achieving the desired objective of reducing pollution is much more likely to be determined. Measuring the amounts of different kinds of pollutants in each area would be an even better guide to determining where programs are and are not successful and the nature of the changes being made. Thus, the value of a measuring tool for accountability is a function of the quality and quantity of the information it provides.

TEST DESIGN AND FORMAT

Examination of an extensive review of published educational measures for elementary school pupils (CSE, 1970) suggests that scores on these instruments are far too often influenced by extraneous factors and, thus, they are not sensitive enough for the purposes for which they are used in performance contracts. This is a rather strong statement, but one, however, which is more than supported by a thorough investigation of some of the typical measures used for accountability. As noted above, one of these extraneous factors is the lack of good overlap between test and program objectives and the priorities attached to these objectives. A second kind of extraneous factor is *test design*. In other words, a student's score on a test may be influenced significantly by how the questions are presented to him as well as by his ability to answer the questions themselves. The formats of some measures are so difficult to comprehend, especially for early primary pupils not used to standardized tests, that one wonders whether the test really assesses the ability listed in its title or simply measures test-taking skills per se. Several well known reading tests for first graders, for example, have two columns of items on each page with eight or more items per column, and four choices per item plus the item stem or stimulus (e.g., a picture). Sometimes this complexity is further compounded by small print, poor and ambiguous drawings, and by how the student has to mark his choice. Typical of the last problem is having to mark in very small boxes next to the item or even on separate answer sheets. This might be acceptable for upper elementary pupils, but clearly inappropriate for first graders not used to taking tests, especially those in the target populations for most performance contracts. One example of this response format problem is illustrated by a frequently used first grade reading vocabulary test. In this test, the first grader is presented with a set of eight pictures and four words. His task is to draw a line from a small star in front of each word to the star in the box that has the picture that goes best with this word. Not only is this task complicated for the first grader, its difficulty is increased by the fact that as he draws lines (with his extra thick, first grade pencil) he crosses out parts of words and pictures. Thus, in answering one question, he reduces the legibility of the next question and the possible answers to it.

INSTRUCTIONS

By now, it may appear that the problems of overlap and test design and format in many existing measures are sufficient to invalidate the results obtained by their use; examination of the instructions and directions bolsters that opinion. Jones (1970) and Kennedy (1970) recently examined these characteristics of tests. Their findings are rather impressive but upsetting. They noted, among other things, that the short term memory load and linguistic requirements of the instructions used in measures for early elementary pupils demand far more than what one could reasonably expect from such students. In other words, the student's score on a measure may largely be a function of his ability to understand unusual and difficult syntax and to follow complex directions rather than of his ability to answer the questions themselves. The directions from a popular reading readiness test illustrate this point:

> Turn to this page, page 9, in your booklets. There are some little stories on this page and groups of pictures that go with the stories on this page. In each little story there is one word that you may not know. Read to yourself the story at the top of the first column. (Pause) You may not know the last word. What sound does it begin with? Yes, with the sound of *t*, as in *take* and *Tom*. Now look at the pictures that go with this story: a *tie*, a *boat*, a *top*, and a *car*. One of these pictures tells what the last word is. Could it be *tie*? No. The word *tie* begins with the sound of *t*, but it is not something to play with. Could it be boat? No. The word boat does not begin with the sound of *t*. Is it *top*? Yes, the word *top* begins with the sound of *t*, and a *top* is something to play with. So put an X in the circle below the picture of the *top* to show that you know the last word in the story is *top* . . .
>
> Now you know what to do with each little story. First you read the story to yourself. Then you make an X in the circle below the picture that tells what the last word is. If you do not know the last word in the story, think what sound it begins with and use the pictures to help you decide what the word is. Remember that the word you choose must make sense in the story. When you finish this page, go on and do the next two pages, also. The arrows at the bottom of the pages tell you to go on.

> The STOP sign will tell you where to stop. You will have 18 minutes for this part of the test. Ready, go!

Although it would appear that these instructions would be sufficient to confuse most first graders thoroughly, there are still some three or four additional lines to be read just to find out whether the child can identify the last word as one of four pictures. As Jones points out, this latter task can easily be accomplished with the use of only a single word. Instructions to teachers (or para-professionals) who might be giving such instruments are equally confusing and often quite vague as to what should be done. The instructions for a test given to every first grader as part of a large western state's mandated testing program illustrate the common problems with such directions:

> Now look at the first row of boxes on the page.
>
> Read the word in the arrow. Which box goes best with it?
>
> See how the box with the bird has been marked. The word is *fly*. The picture of the bird goes best with it.
>
> Now look at the second row of boxes on the page.
>
> Read the sentence in the arrow. Then make a big X on the box that goes best with it.
>
> Did you mark the picture of the car? The arrow says *I went for a ride*. The box with the car goes best with it. (Make sure every child marked the box correctly.)

What happens if the students *did not* correctly mark the boxes to these sample items? There are no instructions in the manual telling the teacher what to do. When this situation happens, as it often does among students with marginal test-taking and/or English skills, test administrators must either rely on their own abilities to explain the sample items further or proceed as if the students really did get the sample items correct. In any event, the so called "standardization" of the test has been broken. If this were not enough to make normative data provided by this measure difficult to interpret, the publishers added the following directions: "Allow enough time, in your mind, for all the children to finish the test." It appears that the publishers assumed that there is unlimited class time or that there are no other pressures (either for or against) giving students more or less time to complete the test. It is illustrative at this point to compare the instructions above with Anastasi's (1968) comments regarding standardized tests. "Standardization implies *uniformity of procedures* in administering and scoring tests. If the scores obtained by different individuals are to be comparable, testing conditions must obviously be the same for all." It is apparent, therefore, that many so-called "standardized" tests do not even come close to meeting the requirements of standardization.

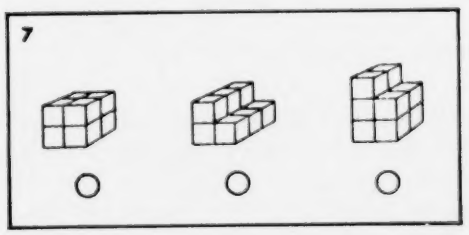
The Center recently completed critical reviews of over 1,500 tests currently being published for students in elementary school (CSE, 1970). In making these reviews, it became apparent that the present crop of standardized measures are often designed much more for administrative ease than for examinee appropriateness. Examples of this include instructions being read only once at the beginning of a test even though half way through the test its format may change, and response formats designed for scoring ease and developed test-taking skills rather than student comprehension of what is required. It is recognized that test publishers must market their products in a profitable manner that is attractive to the purchaser. In the future, hopefully, publishers will pay at least equal attention to the examinee's needs and capabilities so that the scores on their tests can be interpreted directly and not be clouded by irrelevant instructional and format factors.

VALIDITY

So far, three serious limitations of standardized tests in terms of meeting the demands of accountability have been pointed out. These are *overlap* (between test and program objectives and their relative priorities), test *design and format*, and test *instructions*. A fourth problem, and perhaps one of the most serious, is test *validity*. In the typical educational context for accountability, this limitation can be summarized by saying that "the tests do not measure what they purport to measure," i.e., what they measure is not what they might be expected to measure on the basis of their titles or even as they are described in their manuals. To illustrate this point, let us examine the question read to the pupils for an item from a popular reading readiness test:

> Look at the pictures in row 4. Listen carefully: This is a story about three living things you might find around a farm. One day they were talking about how they liked to live. One said, "I like fresh air. When I was very little, I lived in a nest in an apple tree. I lived outside all the time." Another one said, "I like to live outside, too. I lived in a nest when I was very little, but it was on the ground." The other one said, "I don't like it outside very well. I like to live in barns and houses." Find which one spoke *first* and fill in the oval under it. (Note: the three pictures are a mouse or rat, a rabbit, and a bird.)

An analysis of this item reveals that it requires the pupil to store 14 separate units of information, plus sequence, and inferences drawn from the information. At this point, one wonders whether the item belongs on a reading readiness test or a reasoning test or on a listening memory test, but perhaps this issue must await resolution until someone first determines just what prerequisite skills are really needed for reading. The test publisher, however, apparently thought that spatial reasoning and the ability to count were also prerequisites since he included the following item in the same so-called reading readiness test:



Question Read to Pupils
"In row 7 are pictures of three piles of blocks. Each pile has a different number of blocks in it. Fill in the oval under the pile that has 9 blocks in it."

Since the problem of invalid measures goes far beyond reading readiness tests, let us examine some items from popular tests used in performance contracts and state mandated testing programs. For the purposes of illustration, we shall restrict this investigation to first grade reading measures since reading is one of the nation's top priority need areas and the focus of many performance contracts. Exemplary of the items on these measures are the three sets of questions presented below:

> John wanted to buy a cake.
>
> He went to the ___1___.
>
> He also bought some ___2___.
>
> #1 country baker builder airport
>
> #2 butter meat fish bread

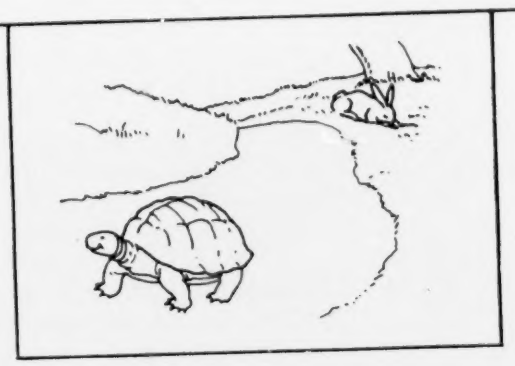
READ THIS A GAME

You say "box"
I say "animal"
You say "fox"

You say "bean"
I say "girl"
You say "Jean"

#3 You say "men"
I say "number"
You say . . . . . four hen ten

#4 You say "head"
I say "food""
You say . . . . . bread cake red



#5

- ☐ The turtle is afraid the rabbit will get ahead of him.
- ☐ The rabbit sleeps while the turtle crawls down the road.
- ☐ The rabbit and the turtle go down the road together.

An inspection of these typical items reveals some interesting common faults. Among the most prevalent of these is that many so-called reading items really measure *reasoning* skills rather than *reading* skills per se. They may be more appropriately included in an intelligence test like Stanford-Binet than in a "reading test." Thus, getting the correct answer on an item is more a function of the pupil's ability to draw inferences from the information presented and, perhaps, previous knowledge than it is his ability to comprehend the meaning of the written word(s). Although some test publishers openly admit that their tests contain such items, others do not. Further, even the publishers who do state in their manuals that the tests contain such items ignore the necessity of providing separate scores for such things as reading comprehension, word recognition, and reasoning. Thus, the test user who thinks he has obtained reading scores is grossly misled. Finally, word-attack skill, rather than comprehension or reasoning, is the primary focus of most first-grade reading programs. In short, the tests are not relevant to the specific aspects of reading that schools expect the tests to measure.

In all fairness to the test publishers, it must be noted that performance on different educational skills are often highly related. Students who have high reasoning ability also generally tend to do well in reading, mathematics, and other school subjects. It is not surprising, therefore, that items assessing different skills are lumped into a single test with only one score provided because the item statistics may give the indication that the items are measuring the same "thing." Since most of the test's items appear to require reading, the publishers further assume that this "thing" is reading. An equally plausible assumption is that two related skills, reading and reasoning, are being measured, and that the reasoning items also require some reading ability. Thus, in order to solve the reasoning items the student must be able to read. Including reasoning items further helps the publisher of standardized tests meet the statistical requirements of "spreading the students out" and having high internal test reliability. A review of typical reading instruments adequately illustrates this phenomenon in that most of the reasoning type items appear toward the end of the test. In short, certain artifactual item statistics may give the publisher a false sense of confidence that his test is really measuring reading when, indeed, it assesses other factors as well.

One undesirable consequence of the foregoing situation is that programs aimed specifically at improving reading (and not reasoning) may be successful in achieving their objectives, but the standardized tests used to monitor their effects may not indicate any significant gains in student performance. As noted earlier, the reason for this is that the standardized test used may not be sensitive enough to reading skills per se since a student's score is due to so many factors. These extraneous factors are not limited simply to problems with format and instructions or irrelevant content, however, but also include the context in which the test is given (such as the room conditions and who administers it), the children's previous experience with standardized tests, and related issues. Items 3 and 4 in the previous example, for instance, actually appear about three-quarters of the way through the test in which they are printed. A first grader having difficulty with these items (and most of them do since the manual indicates that the alternatives are chosen randomly by first graders) may either skip them and go on to the next (and easier) questions or spend a lot of time trying to figure out the answers. Thus, "perseverance" is not an efficient test-taking style on this measure but it does influence a student's score. Similarly, the administration instructions on a well known sentence reading test for first graders changes its avowed focus from reading comprehension to reading speed. It does this simply by requiring the student to answer 15 items like the one below in only five minutes:

Father painted the . . . . . fence fisn front

The preceding discussion considered the first of two accountability demands on standardized tests, namely, providing a valid system for measuring student performance. The examples presented above typify the problems of such measures in meeting this demand, especially in terms of overlap, format, instructions, and validity. In brief, many of the tests do not really measure accurately what they are supposed to measure since the scores they provide are so influenced by a whole host of biases, artifacts, and irrelevant factors. It is not surprising, therefore, that when teachers and principals are held responsible for student performance on such tests that those informed educators among them will have a rather negative attitude toward accountability practices even if they agree with its basic principles. Would a teacher want to be held accountable for student performance on a measure that does not assess the particular objectives with which he is most concerned?

IMPLICATIONS

When school personnel or contractors are held account-

able for their actions on the basis of inappropriate measures, however, we should expect some rather startling results. For example, in 1969 there were over 100 schools in a major metropolitan school district whose first graders were below the 10th percentile in average reading scores. The 1970 test results, however, indicated that there were less than 10 schools below the 10th percentile and that some of the schools jumped more than 50 percentile points. This might be a rather impressive gain were it not for the possible contamination of results due to "special word lists," copies of the test being circulated prior to the formal test administration, and some teachers providing that extra little bit of encouragement and aid during the testing session although such aid is not specified in the test manual.[2] Before condemning these practices, however, we must consider the plight of a teacher or principal in this district who knows that his performance is going to be judged by inappropriate standardized tests and that these results will be published in the local newspapers![3]

Unfortunately, it is the student who loses most by invalid testing techniques. If inappropriate tests and/or testing procedures are employed, then one cannot obtain good information about the relative needs of pupils or the relative merits of different educational programs for meeting these needs. When this happens, one cannot discover which programs would be best for the pupils. Sudden but artificial increases in test scores may also deny schools the very financial aid that they need in order to get real rather than apparent improvements. Further, parents in communities where this is happening know their children are not reading any better and may interpret it as another case of the establishment finding ways to offer inferior educational opportunities; we have all witnessed the effects of such societal attitudes in our daily newspaper headlines.

FAIR PAYMENT SYSTEM

A second implication of the use of inappropriate measures in the administration of performance contracts is that they eliminate the possibility of meeting the accountability demand of providing the basis for a *fair payment system*. In other words, payment is based not on whether the contractor taught students to achieve the desired goals, but is based rather in terms of some extraneous factor such as test-taking skills. When payment is based on such irrelevant factors, the contractor has the choice of teaching what the school feels the student should be learning or teaching them what they have to know to pass a particular test. What the most likely choice will be is obvious especially when there is a lot of money involved.

The appropriateness of standardized tests as the basis for an accountability system also poses some very important technical considerations. It is common practice, for example, to use changes in grade norms between pre- and posttests as the benchmark against which progress is assessed. There are several major difficulties with this approach. First, grade norms are not really what they appear to be. When someone says that a student can read at a grade norm of 3.7 one would assume that he can read as well as the average student in the country who is at that grade level. Unfortunately, this may not be so since this average, generally, is never really computed directly, but is rather interpolated from tests given at the beginning and end of each grade. In other words, the interpolation process assumes that all students progress at the same speed through a grade and learn about the same amount each month. Grade norms also imply that the averages listed in the norm tables are based on a national sample. To support this, the manual might even contain an impressive list of all the schools that cooperated in the norming. What they do not list, however, are all the schools that were contacted but for some reason refused to cooperate. Such refusals are more likely to be systematic and come from low income urban areas than from middle-class white suburbs. It is evident what this bias does to the so-called norms that are actually listed.

A second problem with using score changes between pre- and posttesting as the basis of accountability is that change scores are notoriously poor indicators of student progress (Cronbach & Furby, 1970). The reasons for this are rather complex and cannot be fully discussed in this paper. Some of the practical implications of the problem should, however, be mentioned. Among these are "cellar and ceiling effects" and "regression." The first of these refers to tests that are either so difficult that they do not measure accurately at the bottom end of the performance continuum or so easy that they fail to measure at the top end. If a test has a low ceiling, for example, an average student can gain much more than a bright student. This happens because there is not enough room left at the top of the score distribution for the bright student to demonstrate his increased knowledge or ability. "Regression," on the other hand, refers to the tendency for low-scoring students to improve and high-scoring students to decline between pre- and posttesting solely by chance. Without going into the statistical reasons for this phenomenon, it is apparent that if the students with the lowest scores on a test are selected and then retested sometime later, "improvement" will occur solely by chance. The implications for interpreting the results obtained in many remedial reading programs are obvious. It is apparent, therefore, that the many difficulties with computing change scores warrant extreme caution should they be incorporated into a performance contract. It should also be remembered that the general problems associated with using national norms and/or change scores for accountability persist whether one computes raw test scores, or grade equivalents, or percentiles, or even standard scores.

A third general problem associated with using standardized test scores as the basis of a supposedly fair payment system is that the contract may only call for changes in average scores rather than changes in individual scores. It is obvious, of course, that if all the students in a special reading program improve, then the group's average will also improve. What is not so obvious, however, is that the group's average may go up at the expense of several students doing very poorly (Lindman, 1968). Suppose, for example, that a contractor offered a special mathematics program and a school was going to pay him on the basis of a change in the average score of all the students enrolled in the program. One thing the contractor might do is test the students, identify which ones had the greatest potential for improvement, and then focus all his instructional efforts on them. Though the

[2]*Two kinds of behaviors are covered by the term "Teaching to the Test." The first refers to teaching students to improve performance on the particular objectives covered by the test. This is a good practice if the test covers the objectives you want to achieve. A second definition is teaching the specific items included in a given test. This is not a legitimate practice for the tests used in performance contracts or other accountability situations.*

[3]*The invalid results obtained in the Texarkana performance contract project as a result of the contractor coaching students on specific test items has also indicated definite need for independent audits of evaluation results when such data are used in performance contracts (Welsh, 1970).*

effect of such practices is to raise the group's average, it fails to educate and wastes the time of many pupils in that group.

The several criticisms just noted regarding supposedly fair payment systems are not, of course, aspects of standardized tests per se. They are mentioned, however, because many accountability systems often rely on standardized tests. There are a few instances, however, where some standardized tests may be appropriate for the demands of accountability. For example, suppose a contractor offered to help a high school increase the number of its students being accepted to college. Since College Board Scores are certainly a factor in college admissions, it would be quite reasonable to write a contract in which the scores on the College Board Examination served as one of the criteria of success. In other words, the fact that a test is standardized or used widely, or even if it only provides a single score, does not determine its utility for accountability. The key question is whether the test really measures the relevant performance. Thus, any test that provides useful information about the nature of relevant student performance can be useful for accountability. Unfortunately, almost no standardized tests on the market today meet these requirements for the specific needs of most educational programs for which accountability procedures are being used. Further, this problem is compounded by using with these tests inappropriate benchmarks like national norms and grade-level equivalents as the basis for assessing student programs and paying performance contractors.

RECOMMENDATIONS

So far, the discussion has painted a pretty bleak picture regarding the utility of standardized tests for accountability. The major problems involve questionable test *validity*, poor *overlap* between program and test objectives, inappropriate test *instructions* and directions, and confusing test *designs* and formats. In short, a *VOID* exists between the demands of accountability and the present stock of standardized instruments. Further, this void will probably only widen as the pressure for accountability increases unless we start improving the methods of test construction and use. Among the more important of these improvements would be to provide scores on sets of items measuring a given goal rather than a single global score across whole groups of goals. In this way, we can determine how well students are doing on the particular goals we wish to measure. A second improvement would be to provide more appropriate means of interpreting the test scores obtained without resorting to irrelevant national norms or questional grade equivalents. A third improvement would be to develop test instructions, formats, and designs that coincide with the capabilities and skills of the person taking the test. A fourth improvement would be to tighten test security and administration procedures when test scores are being used for accountability purposes. Finally there should be an increased effort made to develop appropriate assessment methods for important goal areas that are not being measured (or measured well) by existing instruments, especially for high-level cognitive processes and the affective domain.[4] After we have done all of these things, perhaps a brighter picture will emerge when we address ourselves again to the topic of "the uses and limitations of standardized tests in meeting the demands for accountability."

[4]*The Center's projects have addressed themselves to these issues by trying to bring about the desired changes in test construction, administration, and interpretation practices. (Alkin, 1970).*

REFERENCES

Alkin, M. C. Products for improving educational evaluation. *Evaluation Comment*, 1970, 2(3), 1-15.

Anastasi, A. *Psychological testing.* (3rd ed.) New York: MacMillan Company, 1968.

Cronbach, L. J., & Furby, L. How we should measure "change" – or should we? *Psychological Bulletin*, 1970, 71, 68-80.

CSE Elementary School Test Evaluations, Ralph Hoepfner (Ed.) Center for the Study of Evaluation, UCLA Graduate School of Education, Los Angeles, 1970.

Jones, M. H. The unintentional memory load in tests for young children. CSE Report No. 57, May 1970.

Kennedy, G. The language of tests for young children. CSE Working Paper No. 7, February 1970.

Klein, S. P. Evaluating tests in terms of the information they provide. *Evaluation Comment*, 1970, 2(2), 1-6.

Lindman, E. L. Net-shift analysis for comparing distributions of test scores. CSE Working Paper No. 5, March 1968.

Office of Economic Opportunity, RFP No. PRE/E 71-7; July 16, 1970.

U.S. News & World Report, October 12, 1970, 41.

Welsh, J. D.C. perspectives on performance contracting. *Educational Researcher*, 1970, 21, 1-3.

TESTS CITED

Since the purpose of this paper is to note general problems rather than criticize particular instruments, no specific references are attached to the various examples noted. The measures from which these examples were taken, however, are listed below.

Cooperative Tests. Primary. Reading and Mathematics (grades 1-3). Educational Testing Service, 1965.

Gates-MacGinitie Reading Test (Kindergarten). Teachers College Press, 1962.

Metropolitan Achievement Tests. Primary I Battery. Harcourt, Brace & World, Inc., 1958.

Primary Reading Profiles. Houghton Mifflin, 1967 edition.

Pupil Progress Series. Diagnostic Reading. Primary. Scholastic Testing Service, 1956.

Silent Reading Diagnostic Tests. Meredith Corporation, 1970.

SRA Achievement Series. Reading 1-2. Science Research Associates, 1958.

Stanford Achievement Test. Harcourt, Brace, & World, 1964.

CENTER REPORTS

The following Center Reports have been added to the list of CSE Reports currently on file in the ERIC System. For information on how these reports may be ordered, please write to:

ERIC Document Reproduction Service
The National Cash Register Company
4936 Fairmont Avenue
Bethesda, Maryland 20014

CSE NUMBER	TITLE	ERIC NUMBER
2	Feshbach, N. D. Manual of Individual Difference Variables and Measures.	ED 036 877
7	Hagen, J. Program Budgeting.	ED 038 742
8	Pace, C. R. Evaluation Perspectives: 1968.	ED 037 828
9	Bloom, B. S. Toward a Theory of Testing Which Includes Measurement-Evaluation-Assessment.	ED 036 878
10	Scriven, M. Evaluation as a Main Aim of Science: Comments on Professor Bloom's Paper Entitled 'Toward a Theory of Testing Which Includes Measurement-Evaluation-Assessment.'	ED 037 827
11	Glass, G. V. Comments on Professor Bloom's Paper Entitled 'Toward a Theory of Testing Which Includes Measurement-Evaluation-Assessment.'	ED 037 826
12	Guilford, J. P. Comments on Professor Bloom's Paper Entitled 'Toward a Theory of Testing Which Includes Measurement-Evaluation-Assessment.'	ED 036 871
14	Stake, R. Comments on Professor Glaser's Paper Entitled 'Evaluation of Instruction and Changing Educational Models.'	ED 036 872
15	Lumsdaine, A. A. Comments on Professor Glaser's Paper Entitled 'Evaluation of Instruction and Changing Educational Models.'	ED 036 867
16	Gagné, R. M. Instructional Variables and Learning Outcomes.'	ED 036 866
17	Anderson, R. Comments on Professor Gagné's Paper Entitled 'Instructional Variables and Learning Outcomes.'	ED 036 868
18	Postman, L. Comments on Professor Gagné's Paper Entitled 'Instructional Variables and Learning Outcomes.'	ED 036 873
19	Lortie, D. The Cracked Cake of Educational Custom and Emerging Issues in Evaluation.	ED 036 875
20	Gordon, C. W. Comments on Professor Lortie's Paper Entitled 'The Cracked Cake of Educational Custom and Emerging Issues in Evaluation.'	ED 036 874
21	Gage, N. L. Comments on Professor Lortie's Paper Entitled 'The Cracked Cake of Educational Custom and Emerging Issues in Evaluation.'	ED 036 876
48	Sorenson, G., & Hawkins, R. K. Three Experimental Modes of Counseling.	ED 036 879
51	Pace, C. R. An Evaluation of Higher Education: Plans and Perspectives.	ED 036 188
56	Patalino, M. Rationale and Use of Content-Relevant Achievement Tests for the Evaluation of Instructional Programs.	ED 041 044
57	Jones, M. H. The Unintentional Memory Load in Tests for Young Children.	ED 041 043
58	Hoepfner, R. Measuring the Normal Affective States in Children.	ED 039 622

A complete list of CSE Reports may be obtained by writing directly to the Center.

THE CENTER'S CHANGING EVALUATION MODEL[1]

Stephen Klein, Gary Fenstermacher, and Marvin C. Alkin

One of the Center's major purposes is to develop and improve evaluation "theory." In pursuit of this goal, the Center has made a number of changes in its definition and conception of what is involved in an evaluation (e.g., Alkin, 1967a, 1967b, 1969, 1970; Sorenson, 1968; Wittrock, 1966, 1969). This evolution has been stimulated by the writings of our staff and that of other evaluation theorists such as Stufflebeam (1969) and Provus (1969); by an increased knowledge about evaluation studies stemming from our own and others' experiences in actually conducting evaluations; and from the advice of current and potential users of evaluation information (especially those who have participated in our evaluation training workshops). Thus, the Center's model is not fixed (although it often sounds that way when we write about it), but flexible and amenable to change as we learn more about evaluation.

AN INVITATION

In order to develop and improve our model further, we would like your advice. Specifically, we want both to clarify the nature of evaluation and to bring the model closer to describing what is actually involved in a good evaluation study.

The Center believes that the worth of an evaluation model should be gauged by the extent to which it leads to improving educational programs in addition to making summative decisions as to their general worth. In other words, a good evaluation model is one that provides valuable and timely evaluation information to the decision makers who use it. It is apparent, therefore, that a model must guide evaluators in determining what kinds of evaluation information should be reported to decision makers.

Since the readers of *Comment* often function as evaluators and/or users of evaluation information we would appreciate your reviewing and sending us your reactions to our current model. It is hoped that through your comments and criticisms we can improve the model so that it in turn will help to improve evaluation practices.

CURRENT MODEL

Definition. The current model defines *educational evaluation* as the process of *determining the kinds of decisions that have to be made; selecting, collecting, and analyzing information* needed in making these decisions; and then *reporting this information to appropriate decision makers.* Thus, evaluation information should help decision makers in deciding among alternative courses of action, such as how a program might be improved. As may be seen from the figure on the next page, the Center has identified four major kinds of decisions that have to be made (indicated by diamonds) and these are associated with five phases of evaluation activities. The basic features of these decisions and phases are as follows:

Needs Assessment involves stating potential educational goals or objectives (preferably in terms of student performance rather than instructional processes), deciding which of these are of highest priority, and determining how well the existing educational program is meeting these objectives. This latter information is then used by the decision maker to identify the major needs so that he can decide which ones should be attacked. A school superintendent, for example, might have a needs assessment conducted in his district to help him decide where educational programs should be developed or improved. For instance, it might be found that the students at one school are not doing as well as they should in chemistry while at another school the major deficiencies might be in foreign languages. It might also be disclosed that improvements are needed in student performance in English throughout the district. Thus, needs assessment findings are used in determining which problem areas should be attacked.

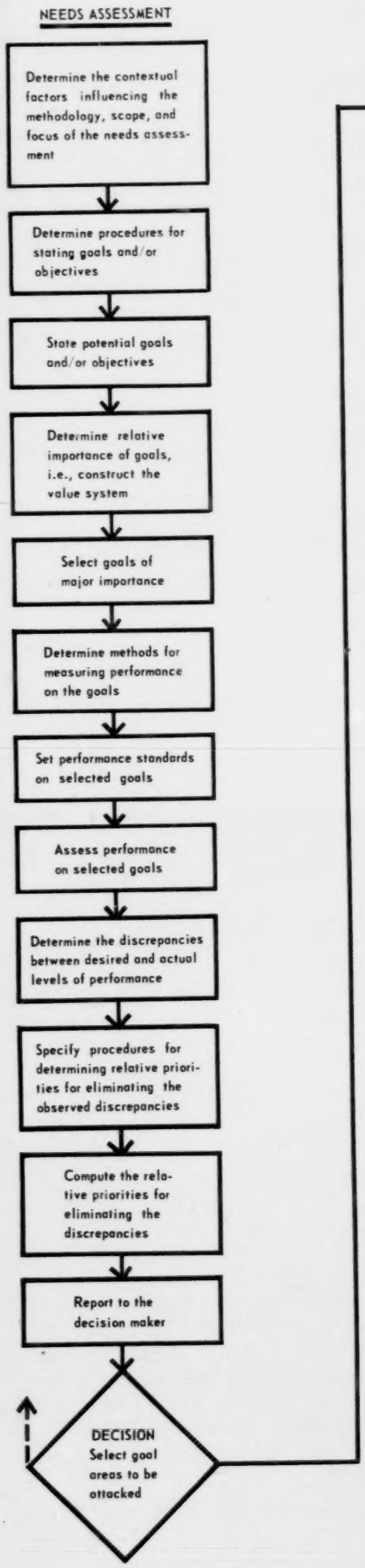
Program Planning involves making decisions about the kinds of programs or combinations of programs (or program components) that should be adopted to meet the problems identified in the needs assessment. Thus, a series of decisions are made about how the needs might best be met with the resources available to do the job. This activity usually involves a series of planning meetings that should result in a written document describing how the school or project intends to achieve the desired objectives. During the program planning phase, the evaluator suggests techniques to facilitate planning decisions, provides advice regarding evaluation requirements for alternative plans, and builds into the final plan the procedures necessary for carrying out subsequent evaluation activities.

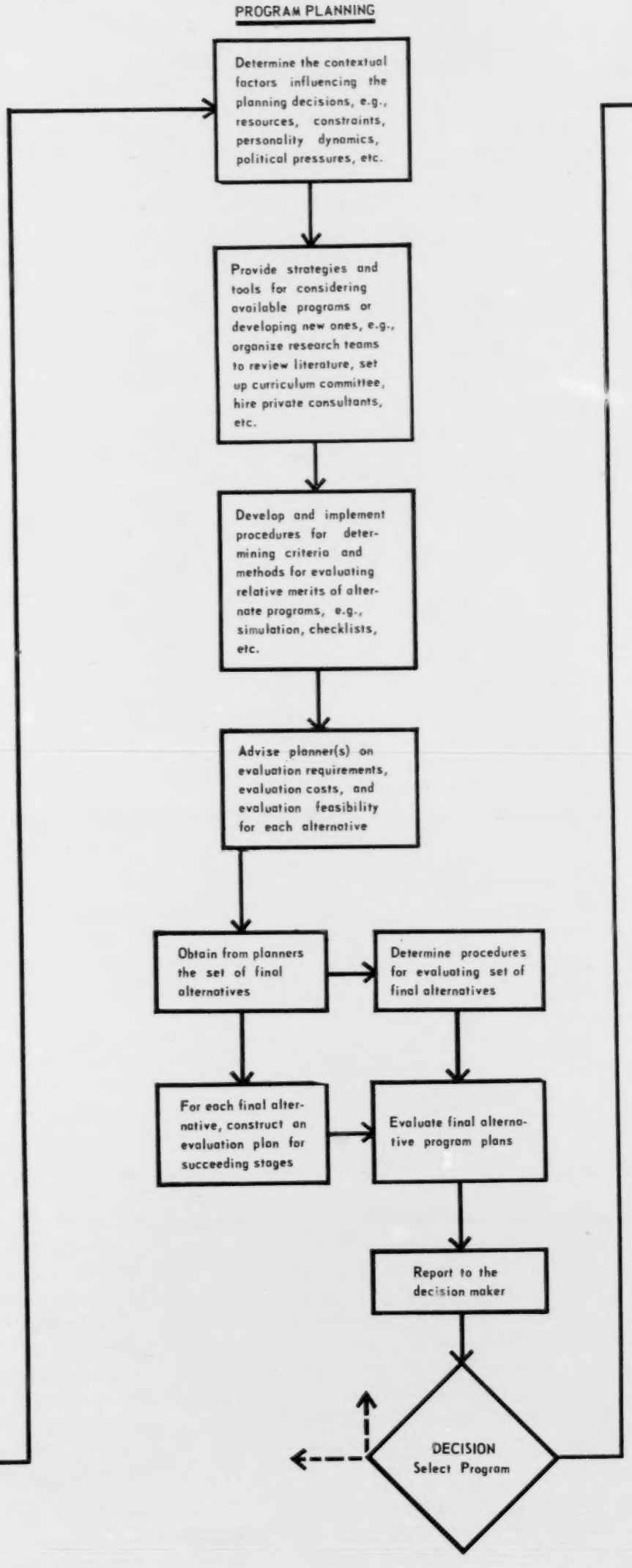
Implementation Evaluation focuses on whether the procedures specified in the program plan are actually carried out in the intended manner. Thus, it involves investigating the degree to which the program plan has been *adapted* properly to the field situation. Typical implementation questions for which evaluation information is needed might be "Did the books arrive on time?" and "Are the students enrolled in the program the ones for whom it was intended?"

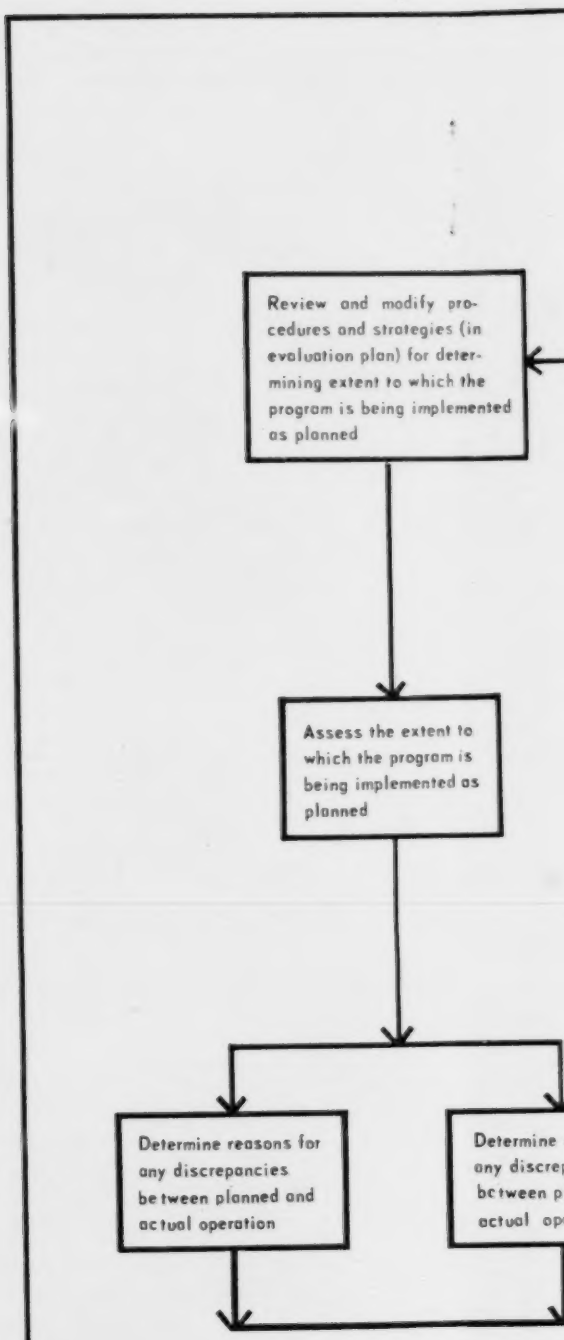
Progress Evaluation, on the other hand, is aimed at determining the extent to which the program is actually making gains towards achieving its objectives. Since a program may be implemented exactly as planned but still not reach its intended objectives, it is necessary to investigate whether the plan is really a good one to achieve the student needs. Further, it is obviously wasteful to install a program in a school in the Fall and then wait until Spring to learn that it failed or that it might have been improved if corrective action had been taken earlier. It is apparent, therefore, that decision makers need information about student progress during the course of a program so that if problems develop they can be identified and corrected quickly.

At this point, it is important to note certain similarities

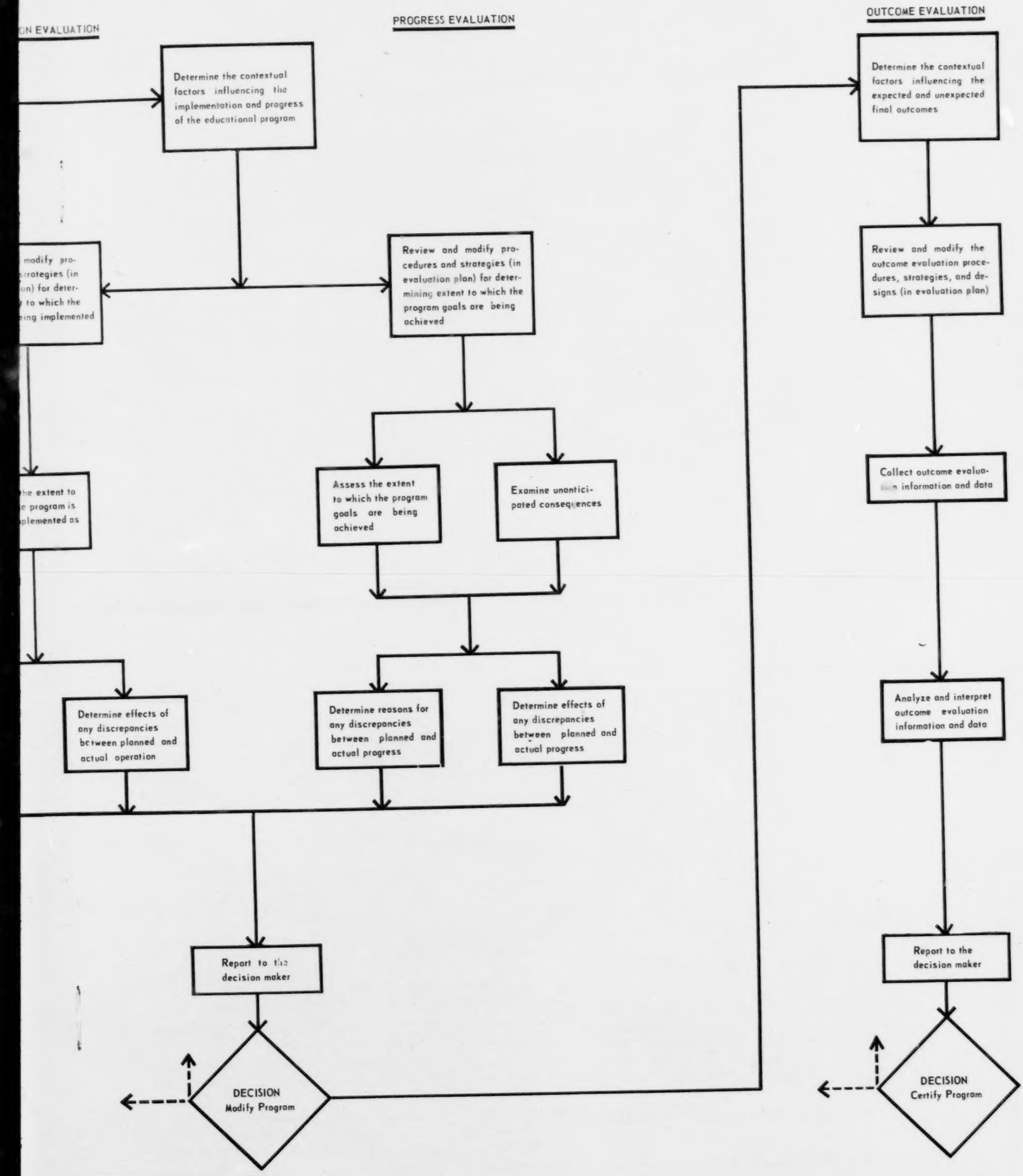
[1]*The major contributors to the development of the Center's evaluation model are Marvin C. Alkin, Gary Fenstermacher, Stephen Klein, Allen Rosenstein, and Rodney Skager.*


NEEDS ASSESSMENT
Determine the contextual factors influencing the methodology, scope, and focus of the needs assessment
Determine procedures for stating goals and/or objectives
State potential goals and/or objectives
Determine relative importance of goals, i.e., construct the value system
Select goals of major importance
Determine methods for measuring performance on the goals
Set performance standards on selected goals
Assess performance on selected goals
Determine the discrepancies between desired and actual levels of performance
Specify procedures for determining relative priorities for eliminating the observed discrepancies
Compute the relative priorities for eliminating the discrepancies
Report to the decision maker
DECISION
Select goal areas to be attacked


PROGRAM PLANNING
Determine the contextual factors influencing the planning decisions, e.g., resources, constraints, personality dynamics, political pressures, etc.
Provide strategies and tools for considering available programs or developing new ones, e.g., organize research teams to review literature, set up curriculum committee, hire private consultants, etc.
Develop and implement procedures for determining criteria and methods for evaluating relative merits of alternate programs, e.g., simulation, checklists, etc.
Advise planner(s) on evaluation requirements, evaluation costs, and evaluation feasibility for each alternative
Obtain from planners the set of final alternatives
Determine procedures for evaluating set of final alternatives
For each final alternative, construct an evaluation plan for succeeding stages
Evaluate final alternative program plans
Report to the decision maker
DECISION
Select Program


IMPLEMENTATION EVALUATION
Review and modify procedures and strategies (in evaluation plan) for determining extent to which the program is being implemented as planned
Assess the extent to which the program is being implemented as planned
Determine reasons for any discrepancies between planned and actual operation

CSE EVALUATION FRAMEWORK


ON EVALUATION
modify pro-
strategies (in
un) for deter-
to which the
ing implemented
the extent to
e program is
plemented as
PROGRESS EVALUATION
Determine the contextual factors influencing the implementation and progress of the educational program
Review and modify procedures and strategies (in evaluation plan) for determining extent to which the program goals are being achieved
Assess the extent to which the program goals are being achieved
Examine unanticipated consequences
Determine effects of any discrepancies between planned and actual operation
Determine reasons for any discrepancies between planned and actual progress
Determine effects of any discrepancies between planned and actual progress
Report to the decision maker
DECISION
Modify Program
OUTCOME EVALUATION
Determine the contextual factors influencing the expected and unexpected final outcomes
Review and modify the outcome evaluation procedures, strategies, and designs (in evaluation plan)
Collect outcome evaluation information and data
Analyze and interpret outcome evaluation information and data
Report to the decision maker
DECISION
Certify Program

and differences between implementation and progress evaluations. Both kinds of activities fall under the headings of "process" or "formative" evaluations and deal with the extent to which the program is functioning properly. Further, both may lead to decisions regarding possible changes and modifications in how the program is being run. Implementation evaluations, however, deal with the extent to which the program's procedures are implemented as planned, whereas progress evaluations are aimed at determining the extent to which these procedures are producing the desired gains in student performance. Decisions about modifying the program will, of course, rely on both kinds of data since there may be problems in how the program plan is being implemented as well as in the plan itself. A more detailed discussion of implementation and progress evaluation issues is presented by Dr. Garth Sorenson in this issue of the *Comment*.

Outcome Evaluations lead to final judgments regarding the general worth of a total program (as opposed to progress evaluations that deal mainly with program components and are done continuously throughout the program's life). Thus, outcome evaluation information is used in making decisions such as "Should we continue the program next year?" and "Should we extend the program to other schools in the district?"

Recycling Loops. The Center's model presents the five kinds of evaluation activities in a logical sequence corresponding to the usual development and operation of an educational program. It is apparent, however, that some of these activities, especially implementation and progress evaluations, may be overlapping in time. It is also apparent that decisions made at one point in a program may require repeating one or more of the preceding phases. A progress evaluation, for example, might indicate poor student performance on certain objectives. A special needs assessment might then disclose that the students did not have the requisite reading skills for the instructional materials specified in the program plan that was adopted to help them achieve these objectives and, therefore, additional planning is needed. To avoid clutter, all these recycling loops have been deleted from the figure, but are implied by the dotted arrows stemming from each of the major decisions. If all the recycling and feedback loops that might potentially be necessary were included in the figure, there would essentially be a line from each box to every other one.

Model Consistency. One important feature of the model is that it has certain consistencies across the five phases. For example, each phase starts with a context determination. The purpose of this activity in needs assessment is primarily to determine the scope and level of the evaluation (e.g., are we evaluating a school or a particular reading program in that school?). Context determination also includes an investigation of the resources, constraints, social dynamics, political pressures, personalities, and environmental conditions that might influence decisions about how program and evaluation activities should be conducted. The nature and focus of context determinations do, of course, change from phase to phase. In program planning, for example, the evaluator would take into consideration the personality characteristics and biases of the planners whereas in implementation evaluation he may focus on potential environmental constraints that may inhibit the program from being run as planned.

Another important consistency is that the second step in each phase involves "setting up" the procedures that will be used in that phase. In other words, a written plan should be developed along with an explication of the rationale for it. This plan describes how the activities in that phase of the model will be conducted. As noted above, it is always possible to revise plans and procedures through recycling, but it is usually better to start with a clear idea of what you intend to do than to assume that so many changes will occur as to make it not worth the effort.

Finally, it is important to note that all evaluation activities in each phase eventually lead to a report to the decision maker who in turn determines whether to drop the project at that point, recycle, or go on to the next phase.

RSVP

We realize that the foregoing brief explanation of the major facets of our evaluation model glosses over many important points and issues. We would appreciate, however, your writing to us regarding your reactions to its general scope, format, orientation, and content. It is hoped that through your comments and criticisms we can improve the model so that it in turn will help to improve evaluation practices. Please address all correspondence about the model to:

Dr. Stephen Klein, Director
Evaluation Theory and Training Program
Center for the Study of Evaluation, UCLA
Los Angeles, California 90024

REFERENCES

Alkin, M. C. Towards an evaluation model: A systems approach. CSE Report No. 43, University of California, Los Angeles; 1967, ERIC Number ED 014-150. (a).

Alkin, M. C. Evaluating the cost-effectiveness of instructional programs. Paper presented at the Symposium on Problems in the Evaluation of Instruction, University of California, Los Angeles, December 1967. Also published as CSE Report No. 25. ERIC Number ED 031-818. (b).

Alkin, M. C. Evaluation theory development. *Evaluation Comment*, 1969, 2(1), 2-7.

Alkin, M. C. A framework to guide evaluation product development. *Evaluation Comment*, 1970, 2(3), 1-4.

Provus, M. Evaluation of ongoing programs in the public school system. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means; the sixty-eighth yearbook of the National Society for the Study of Education, part II*. Chicago: NSSE, 1969. Pp. 242-283.

Sorenson, G. A new role in education: The evaluator. *Evaluation Comment*, 1968, 1(1), 1-4.

Stufflebeam, D. L. Evaluation as enlightenment for decision making. Columbus: Evaluation Center, Ohio State University, 1969.

Wittrock, M. C. The experiment in research on evaluation of instruction. CSE Report No. 3, University of California, Los Angeles; 1966. ERIC Number ED 012-107.

Wittrock, M. C. The evaluation of instruction. *Evaluation Comment*, 1969, 1(4), 1-7.

EVALUATION FOR THE IMPROVEMENT OF INSTRUCTIONAL PROGRAMS: SOME PRACTICAL STEPS

Garth Sorenson
University of California, Los Angeles

During the past decade a movement has been developing which, if properly supported, could contribute greatly to the improvement of education. In general, the movement is an attempt to develop a sound and cumulative knowledge base for instruction together with an adequate educational technology. It is hoped that these practices will enable each new generation of teachers to acquire from the previous generation a repertoire of effective instructional tools, instead of each teacher having to develop his own instructional skills and techniques.

One of the most important aspects of the movement toward improving instructional procedures is an emerging concept which has been called "evaluation of learning experiences" by Tyler (1950), "evaluation for course improvement" by Cronbach (1963), "formative evaluation" by Scriven (1967), and "implementation and progress evaluation" by Alkin (1970). Some have likened it to the idea of quality control in industry. This concept holds that part of the effort and resources ordinarily expended in developing and using any instructional program, whether lecture series, syllabus, textbook, workshop, or training film, should be devoted to testing out and improving that program, particularly during the course of its development, to ensure that the program will work with a particular group of students. Sophisticated program developers do not expect a program to work very well the first time it is tried, and therefore they see it as part of their task in developing any program to take steps to find out why it is not working, to change it, to try it again, and to continue the process until the finished program is effective for those for whom it was intended.

The concept of formative evaluation has evolved in consonance with the gradual shift in the definition of the "good" teacher. In the traditional concept of the teacher's role, the focus of evaluation made good classroom performance an end in itself. The current concern, however, is with the effect of the teacher's methods upon pupil performance as evidenced not merely by academic achievement tests but even more by "criterion-referenced" tests and unobtrusive measures which provide scores on a number of specific objectives rather than a single global score. Contemporary educators may still want students to learn what is in the textbook, but they have begun to reconsider what is meant by such terms as "to learn" and "to know," and they are trying to develop ways to estimate the kind and amount of learning that has occurred in line with these definitions (Klein, 1970). The assumption here is that all students, given proper directions and incentives, can learn a great deal more than they presently do, regardless of I.Q. or aptitude. Further, if a student does not learn in the class, it does not automatically follow that the failure to learn is because of some defect in the student or in the teacher for that matter. The defect is likely to be in the instructional procedures that were used.

The process of finding out at the *end* of a period of time whether or not an instructional procedure has worked with a particular group of students has been called "summative" or "outcome evaluation." Outcome evaluation is a fairly complex and time-consuming process and borrows a number of concepts and procedures from educational measurement and experimental design. As Cronbach (1963) has pointed out, it often requires a great deal of effort just to show that something did not work. (A director of one large education laboratory has described the evaluator as the person "who brings the bad news to the program developer that he has failed again.") It was for such reasons that Cronbach argued that, given the resources and effort required for evaluation, it would be more useful to direct that effort to improving a particular instructional program – that is, to take the "formative evaluation" tack – rather than merely to answer the question whether or not the program produced statistically significant differences in amounts of learning between students taught by that particular method and those who received either no teaching or were taught by another method.

But how to do formative evaluation? It is one thing to insist that evaluation is needed and another to develop workable procedures. This paper outlines a partial model of formative (or implementation-progress) evaluation consistent with new concepts emerging over the past two decades as a result of various R & D efforts. A number of general principles are proposed, a specific illustration presented, and a checklist provided to serve as a guide for evaluators.

A SUMMARY OF FORMATIVE EVALUATION PRINCIPLES

Principle 1. The purpose of any instructional program is to produce measurable changes in the students for whom it was designed; if these changes do not take place, something may be wrong with the program or how it was implemented.

The kinds of changes to be produced by instructional programs include changes in knowledge, feelings, attitudes, etc. For example, some instructional programs are designed to teach people to speak a foreign language, others to play the piano, others to solve problems in calculus, others to understand philosophical concepts. Or instructional programs might be designed to increase a student's feeling of self confidence, particularly in relation to his school work, or to reduce that form of fear and anxiety sometimes called "school phobia." Still others might have the goal of increasing social responsibility or reducing racial prejudice – if anyone knows how to do that.

Principle 2. For any instructional program, it is essential that the goals of the program – whether they involve knowledge, feelings, or attitudes – be defined in terms of performance, behavior, or actions.

No one ever observes "knowledge," "feelings," "sense of responsibility," or "self confidence" directly. We infer each of these characteristics from what a person does, i.e., from his performance or the products of his performance such as, for example, from something he writes. Therefore, in designing a program aimed at increasing a person's ability to read, to understand, or perhaps to create, it is essential that we specify what actions on the part of the learner are to be observed and who will observe

them, in order to make a judgment as to when learning has occurred. We must devise a set of procedures for getting an accurate record of the learner's performance, or at least a reasonable sample of his performance. For these purposes we will sometimes use achievement tests, and sometimes other methods of making observations. Husek (1969) has suggested that the tests or other measures should meet three criteria: first, they should be related to the objectives of the instructional program; second, they should consist of items which few if any of the students answer correctly at the beginning of the course; third, the items should not depend on special language learned in the course, unless the learning of the language is part of the objectives.

It should be emphasized that while the developer of the program should give considerable attention to the kinds of performances he is trying to produce, he should by no means limit his methods of observation to traditional tests. The concept of unobtrusive measures has received attention for a variety of reasons, and in their book, Webb *et al* (1966) suggest a number of alternative directions in which to look.

Principle 3. Instructional procedures should be designed to fit the pre-stated goals – to teach the students the kinds of performance specified.

Obviously, this principle is not to be understood as recommending that students be coached in the answers to standardized achievement test questions, but it does mean that procedures be included that enable the students to learn the kinds of skills which the test will measure. One way *not* to devise an instructional procedure is to select a training film or textbook or to plan a class discussion without first asking, "What do I want the students to learn from this procedure?"

Principle 4. The program developer should follow a theory or model of instruction.

By a theory of instruction is meant a set of propositions about how people can most effectively be taught, together with specific rules based on these propositions to serve as guidelines in such program development activities as preparing instructions to students, arranging sequences of learning tasks, providing for and properly timing the use of incentives, giving students information about their own performance, etc. One purpose of the instructional theory is to reduce the randomness of program planning – the amount of trial and error spent in program writing – and in its place, to develop a set of rules that will enable us to create new and more effective "generations" of programs with less effort. Useful propositions about instruction are to be found in a number of places. Bruner (1961) has described some of the conditions for "discovery" learning. Ausubel (1968) has made suggestions about the use of "advance organizers." Gagné (1965) has presented ideas about task analysis and learning hierarchies. Examples of specific rules are provided by Popham (1970), Stolurow (1961), and others.

As program developers engage in formative evaluation, they should do so with an eye to revising their instructional model as well as the particular program on which they are working.

Principle 5. Instructional programs should be repeatable.

If someone invents an unusually effective method of teaching Russian, or calculus, or the writing of poetry, it is desirable that the operations which constitute that method be described in sufficient detail that other teachers, willing to put forth a reasonable effort to learn those operations, will be able to apply the method with reasonable accuracy. However, to make an instructional program repeatable, it is usually insufficient merely to provide a precise description of the operations. It is also necessary to train other users to conduct those operations in the way that they were planned.

It follows that for each instructional program the developers should provide a set of training procedures for teaching others to use the program as it was intended to be used.

Principle 6. Instructional programs should be pretested.

In developing any program, steps should be taken to guarantee that the program will produce hoped-for changes in members of the target population. To be effective, evaluation should not be delayed until the program has been completed, but should, as Cronbach (1963) has argued, be a part of the developmental operations so as to avoid waste of time, effort, and money.

It is difficult, if not impossible, to evaluate an entire instructional program at once. But it is feasible to evaluate a program one component at a time. For example, a course of study may consist of a number of lessons, each lesson consisting of several parts, and each part requiring say from five to fifty minutes of student time. The evaluation plan should be designed to take each part in turn.

Essential in the process of formative evaluation is that both the program itself and the procedures for training others to use the program undergo the "try-out cycle." As used here, the term "try-out cycle" refers to the following steps:

(a) The component is presented to a small sample – say six students – of the target population.

(b) Its effects on the students are assessed by means of pre- and posttests and other observations. Cronbach reminds us that at this stage it is more important to focus on student responses to individual items than on total test scores.

(c) The component is revised.

(d) It is tried on a new sample.

(e) The cycle is repeated until the component has become demonstrably effective.

Principle 7. Since any given instructional program will work more effectively with some students than with others, the formative evaluation plan should be designed to obtain information about the characteristics of the students, especially those who did not learn from the program.

Instructional programs should be developed for particular target groups – persons about whom certain kinds of information are available or can be obtained – rather than for people in general. Two major categories of student characteristics are obviously important: their previous learnings, and their patterns of motivation. To illustrate: A lesson in advanced calculus would not be appropriate for people who had not already learned beginning calculus. An instructional program in mathematics to be used with students who do not like mathematics would have to be designed to attract, hold, and teach these unmotivated students and would probably be different from one designed for groups of students who were eager to learn mathematics. It follows that for any instructional program it may be necessary to develop alternative components for particular categories of students.

Principle 8. Formative evaluation requires a particular array of roles, skills, and tools which have not traditionally been employed in developing instructional pro-

grams.

Evaluation should not be confused with the more traditional practice of accreditation, which relies on the impressions of experts. Evaluation requires empirically derived information about the effects both good and bad, expected and unexpected, of the program on the students. For the gathering of these data, a deliberate and continuing program must be planned and a staff must be made available and trained to carry out the necessary procedures.

It is easy to point to examples of effective programs, for example in remedial reading, which teachers for some reason have failed to use, even when the use has been approved by school administrators, school boards, etc. It is easy to find examples of programs that teachers use incorrectly, and it is easy to find examples of programs that work well with some students but not with others. Developing programs tailored for acceptance by particular teachers who use them effectively, and designing these programs with sufficient flexibility and discrimination so as to fit the particular students being taught, calls for a special development team equipped with special skills and using special tools. The need for these development team roles and skills has not been recognized by program developers in general.

In developing an instructional program, it is obvious that questions regarding the "content validity" of the programs be answered. For example, are the concepts presented in this program sound, up-to-date, etc? It is less obvious that a number of other technical questions must also be answered such as:

(a) What could go wrong during the instructional process? At what points in the program is failure most likely to occur?

(b) Who is in a position or can be placed in a position to pick up and feed back clues as to the nature of that failure if it does occur?

(c) What procedures are needed to systematically obtain information from the observers about the nature of the difficulties encountered by the users in learning to apply the program, and by the students in learning what the program intends for them to learn?

For such questions to be answered a formative evaluation approach would take into account considerations like these in setting up a program development team:

(a) The users must be trained. It follows that someone needs to play the trainer role.

(b) During the try-out cycle, the users should be monitored to see if they are using each component as planned. It follows that the program development team must prepare monitoring schedules and include someone to play the monitor role.

(c) The students should be pretested and posttested. Someone must choose or build instruments and administer them.

(d) The trainer, the users, the monitor, and the students should be asked routinely and systematically to note where difficulties occur and should be invited to suggest possible solutions to these difficulties. It follows that someone should be assigned the task of asking questions and recording answers. Someone, perhaps the team as a whole, will need to review these answers and make revisions in the program accordingly.

AN EXAMPLE FROM TEACHER EDUCATION

At the risk of stating what to some will be obvious, I would like to give an example to illustrate how the principles of formative evaluation might be applied by someone developing a course in educational psychology for teacher candidates. It is assumed that such a person will not be developing the course in isolation, but will be a member of a team of instructors who are planning a unified and comprehensive teacher training program and who are working together to evaluate the effectiveness of one another's courses.

One of the most useful and most generally taught concepts in American psychology is the concept of reinforcement. Let us suppose the instructor wanted his students to be able to use this concept. Following the line of reasoning outlined here, some of the early questions to be asked by the team of instructors would include, "What do we want teacher candidates to know about reinforcement? Since this course is part of a professional curriculum, what do we want the candidates to be able to do as a result of having learned the concept? What observations would we make in order to determine whether or not a given candidate had achieved a knowledge of this concept?"

In a traditional course in educational psychology, a student might be judged to have learned the concept of reinforcement when he could define the term as it was defined by a particular psychologist, or perhaps when, on a multiple choice test, he could correctly identify which of several definitions of various psychological concepts fitted the term reinforcement. Such a performance on the part of the student would indicate that some degree of learning had indeed occurred, but whether it would have been enough to enable the student to make profitable use of the concept in teaching is questionable.

A professor of educational psychology who tried to follow the suggestions implicit in this paper and who began with the question, "What kinds of behavior on the part of the student would signify that he has learned what we want him to know about the concept of reinforcement?" might postulate a sequence of performances something like those described below. The professor could then infer that a teacher candidate had learned the concept of reinforcement if he were able to:

(a) write a correct paraphrase of the definition given in the textbook, using words other than those used by the author or the instructor. Ability to paraphrase would indicate that more than sheer rote learning had occurred.

(b) distinguish between correct and incorrect examples of the concept presented, let us say, in written form, perhaps on a multiple-choice test.

(c) give new and correct illustrations of the concept — examples other than those given by the textbook or the instructor.

(d) identify logical implications of the concept for teacher behavior.

(e) having observed a teacher instructing a class, describe in writing the actions taken by the teacher which might reasonably be expected to act as reinforcements to a particular student.

(f) describe the steps to be taken in order to determine if a particular set of teacher actions had been reinforcing.

(g) write a lesson plan which prescribed actions to be taken by a teacher in order to reinforce particular kinds of student behavior, such as behavior patterns which constitute good study skills.

(h) describe in writing a strategy for testing the lesson plan — for determining whether or not the planned teacher actions were reinforcing, specifying what observations would be required, who would make them, etc.

Having decided upon the hierarchy of performances he wanted to teach in his course, the psychology instructor, or preferably the team of instructors, would proceed concurrently to do three things:

Evaluation Comment – Page 15

1. The team would develop a set of performance tests which would then constitute an operational definition of each of the course goals.

2. The team would develop a plan for instructing teacher candidates the course concepts in such a way that they could be expected to pass the performance tests. In developing the instructional plan they would presumably follow a theory of instructions such as that exemplified by Popham (1970).

3. The team would devise a strategy for evaluating their instructional plan. In developing the evaluation strategy they might use a checklist* such as the following:

FORMATIVE EVALUATION CHECKLIST

This checklist is for use by developers of instructional programs that are implemented by teachers, counselors, school administrators, etc. Examples of instructional programs might include lectures, remedial classes, workshops, sensitivity training sessions, counseling interviews, etc.

A. Evaluation Plans: Do project plans specify a strategy for formative evaluation, together with a time schedule?
   1. Have outcome measures been developed prior to, or concurrently with program materials? Are there plans for improving outcome measures?
   2. Does the plan provide for a schedule of try-out cycles for each component?
   3. Does the plan include provision for getting feedback from each member of the evaluation team about observed difficulties and potential solutions to problems?

B. The Development Team: Have the following evaluation roles been defined? Who will play each of these roles? Will different persons play the roles of developer and monitor?
   1. Program or materials developer(s).
   2. Trainer(s) of such potential users as teachers, counselors, school administrators, etc.
   3. Monitors of the activities of the users during training.
   4. "Experimental" users, e.g., the teachers who try out the program during its developmental stage and help to improve it.
   5. Data gatherers and processors, e.g., clerks, programmers, coders of tape recorded protocols, etc.
   6. Small samples of the target population(s), e.g., students who serve as subjects and who provide reactions to the program, as well as taking pre- and posttests.
   7. External observers of members of the target population(s). Sometimes outcome measures will involve performances on measures other than standardized achievement tests. For example, performance tests for student teachers would probably require trained raters.

C. Outcome Measures:
   1. Have outcome measures been developed for each component of the program? What evidence is available regarding the validity and reliability of these measures?
   2. Do the outcome measures provide information about possible "side effects" – unexpected outcomes, both positive and negative, e.g., changes in interests and attitudes – as well as attainment of knowledge and skills?
   3. Do the outcome measures satisfy conventions of validity and reliability? For example: since the outcome measures are the operational definitions of the goals of an instructional program, are they indeed viewed by the developer(s) as fully appropriate? If not, are their limitations explicated?

D. The Instructional Program:
   1. Is the program repeatable? Are the steps to be taken by the user/teacher, counselor, administrator, sensitivity trainer, workshop director, etc., – clearly spelled out, in unambiguous language, and in sufficient detail that they can be understood and followed? Have the instructions been pretested?
   2. Is the instructional program organized into testable units – components requiring 50 minutes or less of student time?

E. Formative Evaluation Materials: In addition to the instructional program, have the following materials been developed, or are they in the process of being developed?
   1. Trainers' manuals, which outline specific steps to be taken in training users.
   2. Users' manuals, which outline specific steps to be taken in conducting a particular instructional program.
   3. Users' performance tests, by which to estimate the degree of competence of a user.
   4. Monitors' manuals, including observation schedules which enable the monitor to compare users' performance with a criterion model and to provide information about the number and kind of errors made by a user during a specified period of time. The schedule should enable the monitor to report whether a user is able to follow directions and to inhibit his tendency to improvise; it should also enable the monitor to note the kinds of difficulties the user encounters so that the directions can be rewritten to anticipate these difficulties.
   5. Questionnaires consisting of open-ended questions such as, "What kind of difficulties were encountered in trying to make the program work?" "What suggestions do you have for changes in the outcome measures, the program, the training procedures, the monitoring?" etc. These questionnaires should be directed by the developers to all other members of the evaluation team, including:
      (a) trainers
      (b) monitors
      (c) users
      (d) members of target population
      (e) observers

This checklist was not derived primarily by a deductive process but emerged out of a series of projects aimed at developing effective and reproducible instructional programs. Some of the projects were conducted by doctoral candidates at UCLA who were attempting to develop instructional models in counseling, including Hawkins (1967), Broadbent (1968), Anderson (1970), and Quinn (1970). One project, headed by Hildebrand of the Colorado State Department of Education, was devoted to the development of an effective strategy for the diffusion of improved educational practices to the small rural schools.

F. Measures of individual differences among members of the target population:
   1. Are data being systematically gathered about those individual differences which are likely to make the program more effective with some students than with others, including measures of the following?
      (a) previous learning
      (b) study skills – ability to attend, read, listen, take notes, follow directions, "psych out" teachers, prepare for examinations, etc.,
      (c) attitudes and beliefs – the feeling that one can learn, can cope; the feeling that it is one's responsibility to try, etc.,
   2. Are users, monitors, and observers encouraged to "intuit" crucial differences in students not measured by existing tests in order to provide clues as to the kinds of measures that need to be developed?

G. Administrative Arrangements:
   1. Funds. Has a part of the program budget been earmarked specifically for formative evaluation?
   2. Personnel. Is there a full evaluation team assigned to each program being developed? From the beginning of the project?
   3. Job descriptions. Is there a complete job description for each of the roles listed under B above?
   4. Working relationships.
      (a) Do developers and other members of the evaluation team work cooperatively, or do they see themselves in competition with one another?
      (b) Do developers and other members of the evaluation team see themselves as having a common goal, i.e., the development of a replicable and demonstrably effective instructional program?
      (c) Is each person responsible to one and only one superordinate?
      (d) Does each person know to whom he is responsible?

CONCLUSION

The major emphasis of this paper is that one of the most important services to be performed by evaluation is the improvement of instructional procedures and programs. To achieve this end, the evaluation should focus upon how well the program's components achieve their objectives within the realistic settings for which they were designed. This emphasis will help ensure that the program will work effectively with the particular group of students for whom it was intended.

The advantages to be accrued from formative (i.e., implementation-progress) evaluation seem clear. Rather than waiting to find out at the end of a program whether it has been successful by running an outcome evaluation, it is more useful to direct that effort toward improving the program by testing and refining it while it is still under development. In addition to the greater economy of effort and time offered by this approach is the increased quality and effectiveness of the instructional program. The outline of principles and the checklist contained in this paper present guidelines to facilitate these kinds of evaluations.

REFERENCES

Alkin, M. C. Products for improving educational evaluation. *Evaluation Comment*, 1970, 2(3).

Anderson, E. C. Promoting career information-seeking through group counselor's cues and reinforcements. Unpublished Doctoral Dissertation, University of California, Los Angeles, 1970.

Ausubel, D. P. *Educational psychology: A cognitive view.* New York: Holt, Rinehart and Winston, 1968.

Broadbent, L. A. The effects of two different belief systems on the perceptions of two experimental modes of counseling. Unpublished Doctoral Dissertation, University of California, Los Angeles, 1968.

Bruner, J. S. *The process of education.* Cambridge, Massachusetts: Harvard University Press, 1961.

Cronbach, L. J. Evaluation for course improvement. *Teachers College Record*, 1963, 64(8).

Gagné, R. M. *The conditions of learning.* New York: Holt, Rinehart and Winston, 1965.

Hawkins, R. K. Comparison of three experimental modes of counseling. Unpublished Doctoral Dissertation, University of California, Los Angeles, 1967.

Husek, T. R. Different kinds of evaluation and their implications for test development. *Evaluation Comment*, 1969, 2(1), 8-10.

Klein, S. P. Evaluating tests in terms of the information they provide. *Evaluation Comment*, 1970, 2(2), 1-6.

Popham, W. J. *The teacher-empiricist.* Los Angeles: Tinnon-Brown, 1970.

Quinn, J. B. The influence of interpersonal perception on the process of change in two experimental modes of counseling. Unpublished Doctoral Dissertation, University of California, Los Angeles, 1970.

Scriven, M. The methodology of evaluation. *Perspective of Curriculum Evaluation.* Chicago: Rand McNally, 1967.

Stolurow, L. M. *Teaching by machine.* Cooperative Research Monograph No. 6, U.S. Department of Health, Education and Welfare. Washington: U.S. Government Printing Office, 1961.

Tyler, R. W. *Basic principles of curriculum instruction.* Chicago: University of Chicago Press, 1950.

Webb, E. J., Campbell, O. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago: Rand McNally, 1966.

"MEAN" ELEMENTARY TEST EVALUATIONS

	EDUCATIONAL OBJECTIVE / TEST NAME		MEASUREMENT VALIDITY		EXAMINEE APPROPRIATENESS						ADMINISTRATIVE USABILITY										NORMED TECHNICAL EXCELLENCE						TOTAL GRADES
					Comprehension		Format				Administration			Scoring	Interpretation												
			Content and Construct	Concurrent and Predictive	Content	Instructions	Visual Principles	Quality of Illustrations	Time and Pacing	Recording Responses	Test Administration	Training of Administr.	Administration Time		Norm Range	Norm Interpretation	Score Conversion	Norm Groups	Score Interpreter	Can Decisions Be Made?	Stability	Internal Consistency	Alternate Form	Replicability	Range of Coverage	Gradation Of Scores	
	Rating Range		0-10	0-5	0-4	0-4	0-2	0-2	0-1	0-2	0-2	0-1	0-1	0-2	0-1	0-1	0-2	0-1	0-1	0-3	0-3	0-3	0-3	0-1	0-3	0-2	Good-Fair-Poor
p. 13	GATES-MACGINITIE READING TESTS - READINESS SKILLS Readiness Total	(TCP)	5	0	3	4	1	1	1	1	1	1	1	2	0	1	1	1	1	2	0	0	0	1	3	2	PFFP
p. 31	COMPREHENSIVE TESTS OF BASIC SKILLS Arithmetic - Concepts	(CTB)	7	2	4	3	0	1	0	2	2	1	1	2	1	1	2	1	1	3	1	2	1	1	2	2	FFGF
p. 38	GATES-MACGINITIE READING TESTS Speed and Accuracy (Speed)	(TCP)	8	1	3	3	1	1	1	1	2	1	1	2	1	1	2	1	1	2	0	0	1	1	1	2	FFGP
p. 38	GATES-MACGINITIE READING TESTS Phrases - Flash Presentation	(TCP)	Test mislabled - see below																								
p. 38	GATES-MACGINITIE READING TESTS Words - Flash Presentation	(TCP)	Test mislabled - see below																								
p. 38	GATES-MCKILLOP READING DIAGNOSTIC TEST Phrases - Flash Presentation	(TCP)	8	0	4	4	1	1	1	2	0	0	1	1	1	1	2	0	0	2	0	0	0	1	3	1	FGFP
p. 38	GATES-MCKILLOP READING DIAGNOSTIC TEST Words - Flash Presentation	(TCP)	8	0	4	4	1	1	1	2	0	0	1	1	1	1	2	0	0	2	0	0	0	1	3	1	FGFP
p. 42	GATES-MACGINITIE READING TESTS Speed and Accuracy (Accuracy)	(TCP)	5	1	3	3	1	0	1	1	2	1	1	2	1	1	1	1	1	2	0	0	2	1	1	2	PFGP
p. 59	COMPREHENSIVE TESTS OF BASIC SKILLS Arithmetic - Applications	(CTB)	7	2	3	4	1	1	0	1	2	1	0	2	1	1	2	1	1	2	1	2	1	1	1	2	FFGF
p. 65	GATES-MACGINITIE READING TESTS Speed and Accuracy (Speed)	(TCP)	7	1	3	4	1	1	1	1	2	1	1	2	1	1	1	0	0	2	0	0	1	1	2	2	FFFP
p. 67	GATES-MACGINITIE READING TESTS Speed and Accuracy (Accuracy)	(TCP)	6	1	3	4	1	1	1	1	2	1	1	2	1	1	1	0	0	2	0	0	1	1	3	2	PFFP
p. 83	COMPREHENSIVE TESTS OF BASIC SKILLS Language Total	(CTB)	8	2	3	4	1	1	1	1	2	1	0	2	1	1	2	1	1	2	2	3	2	1	2	2	FFGG
p. 83	COMPREHENSIVE TESTS OF BASIC SKILLS Language - Spelling	(CTB)	7	1	3	4	1	1	0	2	2	1	1	2	1	1	2	1	1	2	2	2	2	1	2	2	FFGF
p. 84	COMPREHENSIVE TESTS OF BASIC SKILLS Language - Mechanics	(CTB)	8	1	3	3	0	1	0	2	2	1	1	2	1	1	2	1	1	2	0	2	1	1	2	2	FFGF
p. 85	COMPREHENSIVE TESTS OF BASIC SKILLS Language - Expression	(CTB)	8	1	3	3	0	1	0	2	2	1	1	2	1	1	2	1	1	2	0	2	1	1	2	2	FFGF
p. 85	COMPREHENSIVE TESTS OF BASIC SKILLS Study Skills - Using Reference Materials	(CTB)	9	1	3	4	1	1	0	1	2	1	1	2	1	1	2	1	1	2	0	1	1	1	3	2	FFGF
p. 87	COMPREHENSIVE TESTS OF BASIC SKILLS, FORM 3Q Arithmetic Computation	(CTB)	9	1	4	4	1	1	1	2	2	1	1	2	1	1	2	1	1	2	0	3	2	1	2	2	FGGF
p. 87	COMPREHENSIVE TESTS OF BASIC SKILLS, FORM 2Q Arithmetic Total	(CTB)	6	2	4	4	1	1	1	2	2	1	0	2	1	1	2	1	1	2	2	3	2	1	2	2	FGGG
p. 89	COMPREHENSIVE TESTS OF BASIC SKILLS, FORM 2Q Arithmetic Applications	(CTB)	7	2	3	4	1	1	0	1	2	1	0	2	1	1	2	1	1	2	1	2	1	1	1	2	FFGF
p. 89	COMPREHENSIVE TESTS OF BASIC SKILLS, FORM 3Q Arithmetic Applications	(CTB)	8	1	3	4	1	1	0	2	2	1	1	2	1	1	2	1	1	2	0	2	1	1	2	2	FFGF
p. 89	COMPREHENSIVE TESTS OF BASIC SKILLS, FORM 3Q Arithmetic Total	(CTB)	6	2	3	4	1	1	1	1	2	1	0	2	1	1	2	1	1	1	0	3	2	1	2	2	FFGF
p. 94	COMPREHENSIVE TESTS OF BASIC SKILLS Reading Comprehension	(CTB)	7	1	3	4	1	1	0	2	2	1	1	2	1	1	2	1	1	2	2	2	2	1	2	2	FFGF
p. 94	COMPREHENSIVE TESTS OF BASIC SKILLS Reading Total	(CTB)	6	2	3	4	1	1	1	1	2	1	0	2	1	1	2	1	1	1	2	3	2	1	2	2	FFGG
p. 94	GATES-MACGINITIE READING TESTS Speed and Accuracy (Speed)	(TCP)	7	1	3	4	1	1	1	1	2	1	1	2	1	1	1	0	0	2	0	0	1	1	2	2	FFFP
p. 96	COMPREHENSIVE TESTS OF BASIC SKILLS Reading Vocabulary	(CTB)	6	1	3	4	1	1	0	1	2	1	1	2	1	1	2	1	1	2	2	3	2	1	3	2	PFGG
p. 96	GATES-MACGINITIE READING TESTS Speed and Accuracy (Accuracy)	(TCP)	6	1	3	4	1	1	1	1	2	1	1	2	1	1	1	0	0	2	0	0	1	1	3	2	PFFP
p.102	COMPREHENSIVE TESTS OF BASIC SKILLS Study Skills Total	(CTB)	3	2	3	4	1	1	1	1	2	1	1	2	1	1	2	1	1	1	1	3	1	1	2	2	PFGF
p.102	COMPREHENSIVE TESTS OF BASIC SKILLS Study Skills-Graphic	(CTB)	9	1	3	3	1	1	1	1	2	1	1	2	1	1	2	1	1	2	0	2	0	1	2	2	FFGP

STATEMENT OF INTENT

The Center for the Study of Evaluation was founded in June, 1966. It is an educational research and development center sponsored by the U.S. Office of Education under the Cooperative Research Act and is the only federally funded center working exclusively on problems in educational evaluation.

The mission of the Center is to produce new evaluation materials, practices, and knowledge which can be adopted and implemented by educational agencies. Emphasis is placed on developing procedures and methodologies needed in the practical conduct of evaluation studies and on developing generalizable concepts and approaches to evaluation problems that are relevant to different levels of education. The Center is directed by Marvin C. Alkin and is staffed by an interdisciplinary team which includes specialists in education, measurement, sociology, economics, and administration.

Evaluation Comment provides discussion of significant ideas and controversial issues in the study of evaluation of educational systems and programs. A copy of *Evaluation Comment* is distributed free of charge to each scholar, researcher, or practitioner on our mailing list. One to five copies may be obtained free of charge; however, where greater quantities are needed readers are encouraged to reproduce the *Comment* themselves. To be placed on our mailing list or to order, subject to availability, additional copies of *Evaluation Comment*, please write to:

James Burry, Managing Editor
Evaluation Comment
Center for the Study of Evaluation
145 Moore Hall
University of California, Los Angeles
Los Angeles, California 90024

CSE ELEMENTARY SCHOOL TEST EVALUATIONS: ERRATA LIST 1

The Center has recently received several letters from interested people that point out errors in the contents of the *CSE Elementary School Test Evaluations*. These diligent readers have discovered both errors of judgment and errors of typesetting. Consistent with its commitment to provide educators with the best possible information for test selection, the Center is publishing a list of corrected entries for those that were found to be in error. The list of test evaluation corrections appears on page 18 of this *Evaluation Comment*. Below are corrected entries for page 136 of the Test Index.

Gates-MacGinitie Reading Tests
- Comprehension 17, 43, 67, 96
- Speed and Accuracy (Accuracy) 42, 67, 96
- Speed and Accuracy (Speed) 38, 65, 94
- Vocabulary 17, 40, 67, 96

Gates-MacGinitie Reading Tests-Readiness Skills
- Auditory Blending 14
- Auditory Discrimination 14
- Following Directions 12
- Letter Recognition 15
- Listening Comprehension 12
- Readiness Total 13
- Visual Discrimination 15
- Visual Motor Coordination 11
- Word Recognition 15

Gates-McKillop Reading Diagnostic Test
add:
- Phrases-Flash Presentation 38
- Words-Flash Presentation 38

The Center regrets all errors in its test evaluations and hopes that the reader will understand both the possibility of their occurrence and the Center's concern that they be corrected.

CSE ELEMENTARY SCHOOL TEST EVALUATIONS

a critical and objective evaluation of all published assessment, diagnostic, and prognostic instruments for elementary school children.

This book contains a compendium of tests, keyed to educational objectives of elementary education, and evaluated by measurement experts and educators for such characteristics as meaningfulness, examinee appropriateness, administrative usability, and quality of standardization. This "periodic table" of tests and objectives is designed for use by principals and superintendents who do not have technical expertise in educational measurement and evaluation, yet its rigorous treatment will make it of interest to educational evaluators and psychometricians.

1970 $5.00 146 pp.

Dissemination Office: ESTE
Center for the Study of Evaluation
145 Moore Hall
University of California
Los Angeles, California 90024

Gentlemen:

Enclosed is my check (money order, or purchase order), payable to the Regents of the University of California, for _____ copies of the *CSE ELEMENTARY SCHOOL TEST EVALUATIONS* at $5.00 per copy, (California residents add 5% sales tax) postpaid. Please send this order to:

Name: _____

Address: _____

_____

City State Zip Code

CSE Announces Publication of the

CSE ELEMENTARY SCHOOL HIERARCHICAL OBJECTIVES CHARTS

This set of 20 charts, printed on heavy paper of various colors and suitable for display, presents a detailed analysis of the 106 exhaustive goals of elementary education at the levels of goals areas, goals, course objectives, curriculum objectives, and instructional objectives; and leads directly to, but does not include, behavioral objectives. The charts are designed to facilitate articulation between broadly stated goals and behavioral objectives so that teachers and principals can translate policy into productive educational activity.

January, 1971 $12.50 20 Charts, 17" x 22"

Dissemination Office: ESHOC
Center for the Study of Evaluation
145 Moore Hall
University of California
Los Angeles, California 90024

Gentlemen:
Enclosed is my check (money order, or purchase order), payable to the Regents of the University of California, for..... copies of the *CSE Elementary School Hierarchical Objectives Charts* at $12.50 per set, (California residents add 5% sales tax) postpaid. Please send this order to:

Name:_____

Address:_____

City State Zip Code

✡ GPO 979-110

UCLA CSE Evaluation comment
Center for the Study of Evaluation

Center for the Study of Evaluation
145 Moore Hall
University of California
405 Hilgard Avenue
Los Angeles, California 90024

NON-PROFIT ORG.
U.S. POSTAGE
PAID
LOS ANGELES, CALIF.
PERMIT NO. 12378

ERIC CTR FOR TESTS
EDUC TESTING SERVICES
MEASUREMENT AND EVALUATION
PRINCETON NJ 08540